Developing Machine Learning to Address Health Inequalities

By

Joseph Abubakar

Date: 24th March, 2025

Version: 1.0.0

Table of Content

1.0 Background of Study

Healthcare could be completely transformed by machine learning (ML), especially in the early detection of chronic conditions like diabetes. However, health inequities that adversely impact disadvantaged and underprivileged groups frequently undermine the efficacy of these models. These differences, which have their roots in racial, social, regional, and cultural variables, might result in biased algorithms and unbalanced datasets that are unable to produce predictions that are accurate for every group. It is feasible to resolve these gaps and guarantee that ML technologies serve all populations equally by utilizing inclusive data practices, sophisticated bias-reduction strategies, and interdisciplinary collaboration. In addition to being a technological difficulty, this is also a moral requirement in the fight for equitable and decent healthcare. AI has the potential to improve the quality and accessibility of healthcare, but in order to achieve the best results, biases, ethical concerns, and technical and financial obstacles must be addressed.

1.1 Objectives of Study

The objectives of this study are outlined below:

1) To analyze and extract insights on what indicators influence inequality in health care, and why those factors influence health care.
2) Develop a mini-batch data pipeline to extract data for automated re-training of the machine learning model.
3) Experiment, develop, and deploy robust machine learning model that is able to classify health inequalities with reference to the indicators.
4) To identify means of mitigating the health inequality.

2.0 Solution Implementation

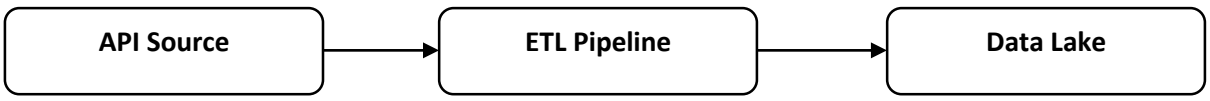
2.1 Data Sources

There are several data sources on the internet. Meanwhile since this project is health related, the data is sensitive due to privacy concern but there are open source datasets that is available for use in projects like this.

1) World Bank Data Repository for Health Inequality: This dataset is available for several indicators such as: health care system and access, mortality and life expectancy, disease incidence, environmental health, etc; Also there is provision for API connection, which means that an automated data extraction pipeline can be developed for access to up-to-date data from the platform. The documentation of the dataset can be found on https://www.who.int/data/inequality-monitor/data and the API documentation can be found on https://www.who.int/data/inequality-monitor/data/hidr-api WHO's Global Health Observatory is also an additional data repository.
2) OpenSafely Platform for Health Inequality Data: This data repository which gives access to over 57 million health records is still under development, including the documentation. The good thing is that they are onboarding people freely to access their platform for analysis and insights. The health data on this repository can be queried using SQL or Python programming language. The documentation can be found here, https://docs.opensafely.org/

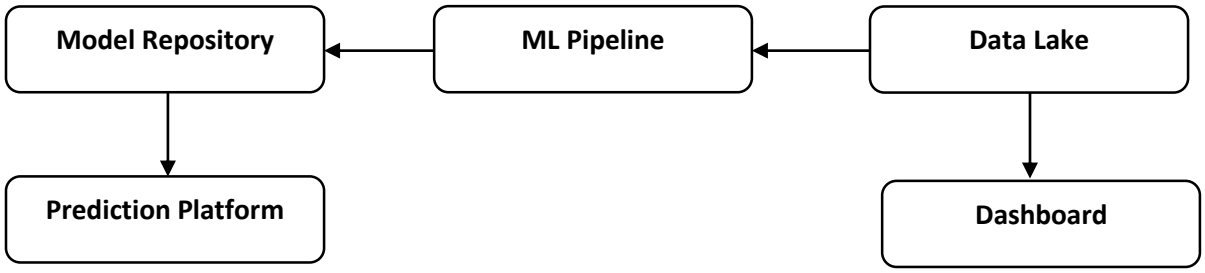
2.2 Data Process

There will be two data pipelines. The ETL(Extract Transform Load) pipeline will extract and store the clean data on Data Lake based on how often the raw data is updated, this pipeline is automated using Github workflow to minimize the cost of infrastructure. Also, extensive cleaning engaged at the transformation stage.



2.3 Data Analytics

The approach is to develop a machine learning pipeline to identify bias or inequality across several indicators, and a dashboard for deeper analysis on the source data to understand why, and how to mitigate this biases. While the machine learning pipeline load the cleaned data, run the experiment over the data, and eject the trained model into model repository. Furthermore, a dashboard will be built to analyze the data store on the Data Lake.



3.0 Conclusion

The biases that exist in AI machine learning procedures in conventional healthcare settings and how they can worsen and prolong current healthcare disparities. In particular, biases in (1) model design, (2) model training and prediction, (3) model deployment, and (4) model evaluation are all investigated. However, the "structural determinants and conditions in which people are born, grow, live, work, and age" are known as the social determinants of health, and they are becoming more widely acknowledged as significant factors influencing health outcomes. Economic stability, the physical and neighborhood environment, education, the community and social context, and the health care system are the five core domains around which Healthy People arranges the social determinants of health.

REFERENCE

Akinyele, Docas & Callahan, David. (2024). Addressing Health Disparities in Machine Learning-Based Diabetes Prediction.

Hosseinpoor AR, Bergen N, Kirkby K, Schlotheuber A, Antiporta DA, Mac Feely S. WHO's health inequality data repository. Bull World Health Organ. 2023 May 1;101(5):298-298A. doi: 10.2471/BLT.23.290004. PMID: 37131942; PMCID: PMC10140685.

Roger Yat-Nork Chung, Ben Freedman. Health inequalities in AI machine learning. Pages 119-130. 2024. https://doi.org/10.1016/B978-0-323-95068-8.00009-1.

S, Madushara & J.B, Buddhinie & Elango, Shopijen. (2025). The Role of AI in Healthcare Inequalities: Can AI reduce or widen healthcare disparities?. 10.13140/RG.2.2.34455.28320.